creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Master's Thesis

Few-shot Domain Adaptation for 3D Human Pose and Shape Estimation

Uyoung Jeong

Department of Computer Science and Engineering
(Computer Science and Engineering)

Ulsan National Institute of Science and Technology

2021

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Few-shot Domain Adaptation for 3D Human Pose and Shape Estimation

Uyoung Jeong

Department of Computer Science and Engineering
(Computer Science and Engineering)

Ulsan National Institute of Science and Technology

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Few-shot Domain Adaptation for 3D Human Pose and Shape Estimation

A thesis submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master

Uyoung Jeong

07/14/2021 of submission
Approved by

_____

Advisor
Kwang In Kim

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Few-shot Domain Adaptation for 3D Human Pose and Shape Estimation

Uyoung Jeong

This certifies that the thesis of Uyoung Jeong is approved.

07.14.2021 of submission

Signature

_____

Advisor: Kwang In Kim

Signature

_____

Kyungdon Joo

Signature

_____

Seungryul Baek

Signature

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Abstract

Despite recent advancements in monocular 3D human pose and shape estimation, many previous works are susceptible to the domain gap between the training data and the test data. This problem become even more severe when the test samples are from challenging in-the-wild scenarios. This paper proposes a domain adaptation approach to mitigate the gap especially in few-shot test environment, utilizing (1) continuous metric loss to constrain the feature space distance relationships between different poses, and (2) segmentation module to localize foreground area so that negative effects from noisy background can be mitigated. Our method achieved slight improvement compared to the baseline on MPI-INF-3DHP and 3DPW datasets.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Contents

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

List of Figures

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

List of Tables

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

I Introduction

3D human pose and shape estimation is a task of regressing 3D human joint or mesh coordinates from input RGB images. It is essential for human body tracking in VR/AR and robotics applications. Recent works significantly advanced the estimation performance thanks to advanced deep neural network architectures, such as fully convolutional network or Transformers. Some other works also tried to exploit temporal information to enhance the prediction results by combining GRU and GAN framework. Still, these models suffer from challenging scenarios, especially occlusion and crowded background. Moreover, the regression performance becomes far more degenerate on unseen cases, due to the domain gap between the train and test data. Training on larger datasets would be one of the solutions, but it is costly to get enough amount of annotations. Some works employed spatial attention mechanism to mitigate the negative effects from noisy background or occlusion, still they do not properly deal with the domain gap problem.

There could be multiple ways to tackle these challenges. If we can enhance the model to directly localize the foreground area, then the negative effects arise from backgrounds could be ignored. In order to provide foreground localization ability, we can either introduce attention mechanism or employ segmentation module.

Another approach is to provide additional regularizer, such as metric learning or viewpoint consistency, in order to better understand the relationships between pose distribution. Though multi-view consistency provides strong supervision, we need to have inputs from multiple cameras, which is quite expensive in reality. We can also incorporate temporal relationship into regularization process, but few works provided a solid solution.

Inspired from these ideas, this paper suggests domain adaptation approach to fill the gap between the challenging test cases and the training data by leveraging metric learning framework and segmentation module. However, its realization could be difficult since attention or segmentation framework also suffer from occlusion, noisy background and crowded scenes. Metric learning loss regularizes the model to distinguish the similar and dissimilar poses in feature space, which can resolve pose ambiguity in challenging cases. While previous metric learning approaches perform time-consuming sampling and loss computation process, we sample and compute the metric loss in mini-batch setting on-the-fly. We exploit segmentation backbone network to explicitly map the person's foreground areas in order to disentangle the backgrounds and the human body. This would help mitigating highly complex background noise. Our contributions are:

- We propose metric loss and triplet sampling method for domain adaptation of 3D human pose and shape estimation.
- We perform multi-task learning with segmentation module to overcome occlusion and noisy background problems.
- Our model achieved superior performance on MPI-INF-3DHP and 3DPW datasets.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

II Related Work

3D human pose estimation is a broad field with various techniques and methodologies. We can roughly categorize the field into several criteria: number of view, stages of network architecture, parametric, and temporality. In this paper, we especially focus on the number of views and temporality. Additionally, we discuss about existing works related to occlusion, attention mechanism, domain adaptation and metric learning techniques.

Multi-view 3D Human Pose and Shape Estimation Iskakov et al. [1] is one of the popular multi-view approaches. Their method performs volumetric aggregation for triangulation from 2D joint heatmaps. Their network is composed of 2D feature extraction model and V2V [2] network. Conventional multi-view methods used RANSAC and Huber loss for triangulation, but it makes the entire model untrainable in end-to-end manner. Instead, they added learnable weights for each camera view and perform aggregation on 3D voxels. Their method was proven to be highly effective on Human3.6M [3] and Panoptic [4] dataset. Xie et al. [5] is another multi-view framework method, which use Panoptic dataset to pretrain the model and fine-tune on the target datasets. They trained the model to generalize on various camera parameters so that the model can adapt to the new camera perspective of the test data. Mitra et al. [6] utilized pose similarity to train the model under semi-supervised learning framework. They exploit cross-view consistency and sample the triplets for loss computation. Note that our method do not use multi-view consistency for metric loss, but we use temporal consistency. Also, their method requires to set the threshold value to distinguish negatives from positives, and tuning the value would become non-trivial. Zhang et al. [7] presented a multi-view model with new indoor dataset. They used limb length and orientation as additional constraints for learning. However, their method requires known camera parameters for projection, and it makes their method unavailable on real-world applications. Tu et al. [8] proposed multi-view and multi-person 3D HPE method, exploiting 3D cuboid proposal network in the middle. Their cuboid proposal network proposes a bounding box on the 3D space, and V2V regression network predicts the pose from the given bounding box. However, its cuboid proposal network would significantly increase the number of learnable parameters, making the model computationally expensive.

Monocular 3D Human Pose and Shape Estimation. We primarily focus on monocular settings, where the models estimates the human pose from a single view. One of the popular approaches is to use a parametric model called SMPL [9] to regress the human shape. SMPL(Skinned Multi Person Linear Model) is a blend skinning function composed of template mesh at rest pose, joint positions, blend weights and angular pose of the skeleton. If we provide proper shape(β) and pose(θ) parameters, regressor of the SMPL produces a human mesh. In order to regress the SMPL parameters, several works [10–13] stacked several fully connected layers on top of the ResNet [14] backbone. One of the limitations of this approach is lack of ability to regress the SMPL parameters correctly. Most of the parameter regression capability is confined within the shallow fully connected layers, and its 2D CNN backbone is not sufficient to capture complex 3D geometry of human articulation. Therefore, these methods often produce

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

unrealistic prediction results.

HEMlets PoSh [15] takes relative joint-wise local depth order to learn depth relationships. They map the relative depth order as -1, 0 or 1. Although they showed notable performance improvement, it would be more informative if we deal with the problem in continuous domain.

VIBE [13] is one of the recent works based on monocular and temporal estimation method using parametric model. VIBE is built upon adversarial learning framework, where pose regressor serves as a generator and the motion discriminator regularizes to produce realistic pose sequence. Although it proved its effectiveness on 3DPW and MPI-INF-3DHP datasets, VIBE fails on unseen data, especially large pose or shape variations, occlusion or multiple person cases. One of the reasons is the ability of the discriminator, since its training data could not cover the large pose variations in wild scenarios. Also, the temporal module consists of simple GRU cells with little engineering, which induces inferior temporal regression capability. Subsequent works [16, 17] also could not fundamentally overcome the above limitations.

Instead of the parametric regression models, several other approaches have also been proposed. Wang et al. [18] proposed a temporal body model estimation model from point clouds with MLP layers. Lin et al. [19] proposed a non-temporal mesh regression model called METRO, which consists of Transformers architecture, which is one of the best-performing models on MPI-INF-3DHP.

Sun et al. [20] proposed a fully convolutional network, called ROMP, for SMPL parameter regression. It ranked the first place on 3DPW dataset at the time of writing. Although its performance is impressive, the model may suffer from multiple local maxima, resulting in suboptimal predictions.

Although METRO and ROMP improved the estimation performance far better, these models still cannot fill the domain gap effectively. Even maintaining a consistent coordinate systems between the datasets is a complex and tedious process, and there is no such rule of thumb for them.

Attention Mechanism. Many researchers have been working on how to deal with the challenging situations of human pose estimation. Attention mechanism is considered as one of such approaches. Chu et al. [21] and Lie et al. [22] adopted attention mechanism by applying learnable weighting on each intermediate feature of the neural network architecture. Liu et al. [23] suggested channel-wise and frame-wise attention module to provide temporal and kinematic understandings. Zhu et al. [24] exploited attention mechanism for pose transfer task. They built a network which transfers the pose of the input image to the target pose, by gradually morphing the pose across the network blocks. Their learning framework is based on GAN architecture, where pose transfer network serves as a generator, while appearance discriminator and shape discriminaotr supervises the overall image-to-image translation process. Chen et al. [25] tackles image captioning problem with spatial and channel-wise attention map. Their method can be interpreted as applying same spatial attention on multiple channels, discarding other channels' information. Li et al. [26] solves semantic segmentation task using attention map with self-guidance. They claimed that since pixel-label annotation is not sufficient, they suggested to use image-level labels to solve the problem. Their network is composed of classification and attention networks, where attention maps are made by Grad-CAM framework. They evaluated on PASCAL VOC

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

2012 [27] and conducted comparative experiment on custom subset. Xu et al. [28] solved depth estimation problem using attention-guided CRF(conditional random field) implemented as neural network. Their attention mechanism constrains the amount of information flow from multiple scale features to the final attention map during fusion process. Fukui et al. [29] focused more on the visual explanation of image classification using attention. Their branched network structure is supervised by conventional cross-entropy loss. Fu et al. [30] applied positional attention and channel attention on the output feature of the CNN and fuse them with element-wise summation. Their ablation study shows performance improvement on Cityscapes [31] validation dataset. Zhang et al. [32] performed visual localization task, which is to predict the location that the picture is taken. Evaluation is conducted by retrieving an closest image to the query. Their network gets point clouds and process through the MLP layers. Sampling and grouping layers act as attention on each points and perform feature accumulation. By iteratively sampling centroid points and feature processing, their representation shows better localization result on Oxford [33] test dataset. Xie et al. [34] claimed that image impainting task has several challenging issues, such as irregular holes and blurred color. They made bidirectional binary mask with hand-crafted convolution kernel. Their method seems heuristic and lacks ablation study about selection of the convolution kernel. Xia et al. [35] solves person re-identification problem. Their method computes covariance matrix from intermediate network features to get second-order correlation information.

Occlusion. Chen et al. [36] solves monocular 3D human pose estimation under partial occlusion scenario. They masked out some part of the keypoints or frames by setting corresponding heatmaps to zero, while spatio-temporal discriminator validates the predicted pose sequence. They additionally added random noise and translation for augmentation. Their method showed better result on MPI-INF-3DHP, 3DPW and Human3.6M datasets. Zhao et al. [37] filters occluded areas using occlusion mask for optical flow estimation. they achieved state-of-the-art on MPI Sintel [38], KITTI 2012 [39] and 2015 datasets. Jiang et al. [40] suggested occlusion-aware indoor 3D scene understanding technique. Their method predicts the occluded part of planes in indoor scenes and perform plane warping. The network is trained by multi-view consistency of planes. In order to provide ground-truth, they generated pseudo labels for training. Kortylewski et al. [41] addressed object classification with partial occlusions. They made a multi-task model composed of occluder kernels and class mixtures, providing localization of partially occluded objects.

Domain Adaptation. Domain adaptation has been actively researched across various tasks in computer vision. It is deeply related with domain transfer and few-show learning, and we may mix these terms in this paper. Piao et al. [42] proposed a face shape estimation method by transferring the input real face images to the synthesized 3D face mesh. Sundermeyer et al. [43] adapted the object pose estimation network from the synthetic to real data. They performed rigorous data augmentation to generalize the model. Wang et al. [18] suggested a weakly supervised learning method for 3D point cloud segmentation. They built a template shape pool and retrieve the template similar to the input point cloud. Then the model deforms the shape towards the input shape while preserving the labels of the template point

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

clouds. Tang et al. [44] utilized the pose estimation results for image classification task. They used normalized pose heatmap as an input of the classification network. Cao et al. [45] suggested animal pose estimation method by utilizing human pose dataset. They constructed two-way network, composed of domain discriminator and domain adaptation network, and then feed the output as an input of the keypoint estimator. They trained the model by pseudo-label-based optimization process.

Many other works focused on domain adaptation on few-shot scenario. Yang et al. [46] tackled the face manipulation detection problem by minimizing the distance between the synthesized image and the Deepfake image on the embedding space. Ling et al. [47] built a network composed of segmentation, contour and texture map networks and recognize the pills from the images using triplet loss. Bateni et al. [48] used Mahalanobis distance with class covariance matrix to perform the few-shot classification problem. Benzine et al. [49] solved multiple human pose estimation task by anchor selection method. Anchors are pre-defined bounding boxes which serve as templates for person detection. Wang et al. [18] Jaritz et al. [50] suggested a two-stream network for point cloud segmentation. One network gets an RGB image and the other gets point clouds and constrain the networks using KL divergence between the projections and the 3D point clouds.

Deep Metric Larning. Distance metric learning has been applied to wide range of computer vision tasks. Its basic idea is to map the data into the embedding space with specific distance relationships based on the similarity of data features. Chopra et al. [51] designed contrastive loss which constrains the model to map the two different input data into the embedding space. If the two inputs are the same class, the model should minimize the distance between them. Otherwise, the distance should be larger than the pre-defined coefficient. Triplet loss [52] is one of the popular metric losses applied to various tasks. It is composed of anchor(query) f, positive f^+ and negatives f^-s. The loss tries to keep the distance between $|f - f^+|$ and $|f - f^-|$ with a certain amount of margin m. Several other losses [53, 54] have been suggested, but these losses cannot be directly applied to our problem, since they are designed for classification task. In order to apply the conventional metric loss into the 3D HPE problem, labels should be discretized. Kim et al. [55] introduced 'Log-ratio loss' using continuous labels. They evaluated their method on pose similarity retrieval task, and showed superior results compared to the triplet loss. We will discuss more about the loss in the section 3.2.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

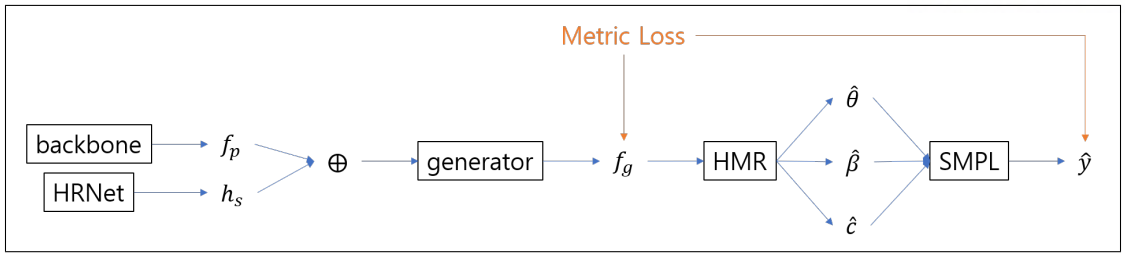


Figure 1: Overall pipeline of our network. We use VIBE architecture as a baseline network. We regularize the generator output and the final regression output using metric loss. Our segmentation module serves as a multi-task backbone network for foreground feature extraction.

III Method

3.1 Problem Overview

Our problem is based on monocular and temporal 3D human pose and shape estimation task. Additionally, we assume few-shot test environment. Our overall pose regression network can be expressed as $\mathscr{F} : x \mapsto \{\beta, \theta, c\}$, where $x \in \mathbb{R}^{W \times H \times 3}$ is an input RGB image, and $\beta \in \mathbb{R}^{10}, \theta \in \mathbb{R}^{24 \times 3}, c \in \mathbb{R}^3$ are shape, pose and camera parameters respectively. β, θ are then used to reconstruct SMPL body mesh, and the joint regressor of the SMPL produces the 3D joint locations $\hat{y} \in \mathbb{R}^{K \times 3}$, where K is the number of keypoints. c can be used to project the 3D joints onto the image plane. As illustrated in Figure 1, our implementation is based on VIBE [13] architecture, and we would not discuss much about the VIBE itself. VIBE pose estimation network is composed of pose feature extraction backbone F_p, pose generator G composed of GRU cells, and SMPL parameter regression network H. Thus, baseline VIBE can also be expressed as $\mathscr{F} = H \circ G \circ F_p$.

On top of the baseline model, we introduce a segmentation module F_s which performs foreground area feature extraction along with the F_p. Then, our model can be expressed as $\mathscr{F} = H \circ G \circ (F_p \odot F_s)$, where $\odot$ is a element-wise fusion operation, such as Hadamard product. In order to fuse the features, however, we need to perform additional nontrivial engineering process, which we will discuss on section 3.6.

3.2 Log-ratio Loss Review

In log-ratio loss, features and labels of anchor a and other samples i, j are used. Feature of the anchor is denoted as f_a, and the corresponding label is y_a. Similarly, features and labels of i, j are referred as f_i, f_j and y_i, y_j. Then, the log-ratio loss is defined as equation 1.

$$\mathscr{L}_{lr}(a,i,j) = \left\{ \log \frac{D(f_a, f_i)}{D(f_a, f_j)} - \log \frac{D(y_a, y_i)}{D(y_a, y_j)} \right\}^2 \tag{1}$$

The loss minimizes the distance between the distances in the feature space and the label space, preserving the label space distance in the embedding space. Despite its effectiveness, it induces insta-

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

bility during training due to high loss fluctuation, since logarithm of [0,1) range produces high negative values, and this is an unexpected adversarial effect of the loss design. Additionally, *NN* size distance matrix should be constructed for loss computation, inducing scalability issue on large data. While Kim et al. [55] trained the model on $\tilde{1}$2K training data, VIBE is trained on $\tilde{5}$9K samples. If the log-ratio loss is directly applied to the VIBE, 59K $\times$ 59K distance matrix should be constructed, where the number of elements exceed 3,481M. Note that we ignored the number of sequence length of each batch sample.

3.3 In-Batch Triplet sampling

In order to compute metric loss in a tractable way, we sample the loss components within general mini-batch setting. In order to sample the anchors, we randomly select one frame per one batch sample. If the batch size is denoted as B, then we get B anchors. In case of positive samples, we select all frames within the sample batch sample, excluding the anchor and temporally nearby frames. We set 'skip step' parameter to define the nearby frames, and it is set to 3 by default. If the sequence length is denoted as seq_len, we get $seq\_len - skip\_step - 1$ positive samples in total. All the other samples are then set as negatives.

In summary, we sample the metric loss components for each anchor as below:

- positive (include nearby frames): $\#(seq\_len - skip\_step \times 2 - 1)$
- positive (exclude nearby frames): $\#skip\_step \times 2$
- negative: $\#(B-1) \times seq\_len$

Compared to the original log-ratio loss, it does not require any additional preprocessing on the dataset, and also does not have scalability issue over the total number of dataset. Our method calculates the loss $O(B^3 \times seq\_len)$ at the worst case, where skip step is set to 0. Since the batch size is significantly smaller than the total data size, it is practically more tractable. One of the limitation is lack of hard negative sample mining, since we cannot effectively perform hard negative mining within the given mini-batch. Also, treating the consecutive frames as positive samples would hinder the estimation performance, when the pose quickly changes over short period of time.

3.4 Metric Loss

Based upon the equation 1, we add an offset 1 and compute the loss using the components defined in the section 3.3.

$$\mathscr{L}_{lr+}(A,I,J) = \frac{w}{|I| \times |J|} \sum_{a}^{A} \sum_{i}^{I} \sum_{j}^{J} \log \left\{ 1 + \left\| \frac{D(f_a, f_i)}{D(f_a, f_j)} \frac{D(y_a, y_j)}{D(y_a, y_i)} \right\|_1 \right\}^p, \tag{2}$$

where A, I, J are set of anchors, positives and negatives. p is the degree of norm, which is set to 1 as default. w scales the loss.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Our loss is numerically more stable, while preserving the idea of continuous distance mapping between the feature space and the label space. Moreover, the positive samples are temporally related, so the model should learn the temporal similarity. Also, we sample multiple anchors from the mini-batch, so that we can further generalize the loss computation. One of the major drawbacks is intensive computation compared to the conventional loss design, since previous works tend to consider only one anchor.

3.5 Metric Loss for Few-Shot Domain Adaptation

In this section, we reformulate the aformentioned metric loss in few-shot domain adaptation setting. We assume that there are source and target datasets, and we try to adapt the model pretrained on the source dataset to the target dataset. We propose the following training pipeline to perform domain adaptation using metric loss.

Algorithm 1 Pseudocode of few-shot domain adaptation with metric learning

for $a \in T$ **do**
  for $s \in S$ **do**
    $f_a = \mathscr{F}(a)$
    $f_s = \mathscr{F}(s)$
    $D_y = L2Dist(y_a, y_s)$
    $D_f = L2Dist(f_a, f_s)$
    sample positives and negatives, where $D(y_a, y_{pos}) < D(y_a, y_{neg})$
    backpropagate the loss computed by $MetricLoss(D(y_a, y_{pos}), D(y_a, y_{neg}), D(f_a, f_{pos}), D(f_a, f_{neg}))$
  end for
end for

In the algorithm 1, S is the source dataset and T is the target dataset, where a and s are the mini-batches. For more advanced sampling, we can sample the nearest n samples as positives, and the farthest n samples as negatives. Note that this metric loss framework is different from the sampling method discussed in the section 3.3, since we do not sample temporally similar frames in this framework. Using this training procedure, we can ensure that the model can learn the similarity and dissimilarity between the source and the target datasets in the feature and the label space. Although its triplet sampling is the same as the original log-ratio loss, its criterion to divide the positives and negatives is too naive. If the samples have no notable difference in terms of pose, the metric loss cannot provide meaningful supervision to the model, which can lead to suboptimal results.

3.6 Segmentation Module

Segmentation module can be interpreted as a foreground mask extractor, which gives a valuable feature for precise human body localization. Instead of applying attention mechanism, producing segmentation map with separate network would be more effective during inference, due to the difference of network

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

capacity. However, the most challenging problem is the lack of pixel-wise segmentation annotation on existing 3D human pose datasets. Since there is no ground-truth annotation, we generate pseudo ground-truths on 3D human pose dataset. After the annotation is prepared, we train the segmentation module by using IoU score between pseudo ground-truth and the predicted map.

We generated pseudo-ground-truth using ensemble of Mask-RCNN [56], Cascade Mask-RCNN [57], and 2D keypoint annotation from MPI-INF-3DHP dataset. First, we make a 2D patches of a skeleton mask from 2D keypoint annotation, which will serve as a basic segmentation mask. Then, we infer separate segmentation masks from each instance segmentation models and perform conjunction operation for each pixel so that we will have only the overlapping areas, since it would be more robust and reliable to be used as a ground-truth. After conjunction operation, we finally get union of the inferred segmentation maps and 2D keypoint masks as a final pseudo-ground-truth. If we denote s' as a pseudo-ground-truth segmentation mask of the input image x, then our overall ground-truth generation process can be defined as equation 3.

$$s' = (\text{Mask\_RCNN}(x) \cap \text{Cascade\_Mask\_RCNN}(x)) \cup \text{2D\_Keypoint\_Mask}(x) \tag{3}$$

Additionally, we filter out the nonnegative values below the threshold, 0.2 by default.

In order to fuse the predicted segmentation maps and human pose features, segmentation maps should be further processed to have same dimension of the human pose feature. Human pose feature $f_p \in \mathbb{R}^e$ is a feature vector with dimension e, so we need to construct a dimensionality reduction function $F_d : f_s \mapsto \mathbb{R}^e$, where $f_s \in \mathbb{R}^{w \times h}$. After the dimension has been reduced, we fuse those features with element-wise multiplication operation. To summarize, our overall feature fusion process can be expressed as equation 5.

$$f = f_p \otimes F_d(f_s) \tag{4}$$

After the fusion, f is fed into the pose generator and subsequent networks, producing predicted human shape at the final stage.

$$\hat{y} = H(G(f)) = H(G(f_p \otimes F_d(f_s))) = H(G(F_p(x) \otimes F_d(F_s(x)))) \tag{5}$$

There are two ways to train the segmentation module. One way is to jointly train the segmentation and pose estimator at the same time, and the other way is to pretrain each network and fine-tune the feature fusion network, without further training the segmentation module. Since our pose estimator is based on GAN architecture, joint training might cause mode collapse, which should be avoided if possible. Therefore, we pretrain the segmentation module and finalize the weights, and then fine-tune the feature fusion network by jointly training with the pose estimation network.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

IV Experiments

4.1 Dataset

MPI-INF-3DHP [58] is a main training dataset in this paper. Video is taken in a indoor studio. Training data contains 8 different subjects, 14 cameras, and two different outfits. *seq_len* is set to 16, which is the same as the VIBE, producing 59K samples. **3DPW** [59] is a wild 3D human pose dataset with 60 video sequences and different clothing variations. We evaluate our method on this dataset without additional training or fine-tuning. Following the VIBE training framework, PennAction [60] is used for 2D keypoint supervision, and AMASS [61] is used to provide ground-truth for the discriminator. We primarily use MPJPE(Mean Per Joint Position Error) and PA-MPJPE as evaluation metrics. PA-MPJPE computes MPJPE from the keypoints aligned by the base keypoint(e.g. pelvis). PVE(Per Vertex Error) and Accel(Acceleration error, mm/s^2) are also reported for further analysis. Acceleration error is the difference of joint locations between adjacent frames.

4.2 Implementation Details

We use the architecture from VIBE [13]. Although several hyperparameter configurations have been tested, but we found that hyperparameter tuning affected little to the performance in overall. We primarily use HRNet-OCR-W48 [62] as our segmentation module. We modified the network output channel size to 2: one for foreground and the other for background classification. In case of feature fusion network, we employ ResNet-50 network pretrained on ImageNet [63] dataset. We resize the input image to 520×520 size. We also tested with different class weight configurations, as we will further discuss on the section 4.4. For few-shot testing, we use MPI-INF-3DHP as source dataset, and 3DPW validation set as target dataset. For triplet sampling, we used 128 batch size. We sample top 30% similar and dissimilar samples as positive and negatives, respectively.

4.3 Evaluation Results: Metric Loss

Table 1 shows the experiment results on 3DPW validation dataset. Compared to the baseline VIBE, our method is slightly inferior in terms of MPJPE and PA-MPJPE, while superior in terms of PVE and Accel. We also conducted experiments with different weighting factor values of the metric loss for ablation study. As w increases, the metric loss more aggressively constrain the model to produce temporally smooth results, thus acceleration error is decreased. One may find it questionable why PVE is improved while MPJPE/PA-MPJPE is degraded. One of the possible explanation is that overall body pose and shapes are temporally has become more consistent while sacrificing the joint location accuracy.

We also conducted ablation study about positive sampling strategy and skip steps. Although sampling nearby frames with 3 skip steps and 1.0 weighting showed superior performance in terms of MPJPE, PVE and acceleration error, excluding nearby frames showed better performance in terms of PVE and acceleration error in overall. Interestingly, increasing skip steps with nearby positive sampling degraded its performance.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Method	positive sampling	skip step	w	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
VIBE	exclude nearby	-	-	93.9613	**60.0187**	118.7732	30.0591
Metric	exclude nearby	3	0.1	**92.4542**	60.7224	116.7338	29.3358
Metric	exclude nearby	3	0.2	93.8174	60.8965	123.2773	31.3126
Metric	exclude nearby	3	0.5	93.3917	60.6369	**113.4797**	**27.1980**
Metric	include nearby	3	0.1	94.8721	61.0294	117.8031	30.4547
Metric	include nearby	3	0.5	96.5314	60.7599	118.7517	28.7588
Metric	include nearby	3	1.0	92.5125	60.1324	118.1193	28.5850
Metric	include nearby	5	0.5	95.7067	61.3537	122.7755	30.4942
Metric	include nearby	5	1.0	97.0412	61.3988	122.1917	30.2956
Seg	-	-	-	94.6754	60.1173	118.4941	33.6988

Table 1: Evaluation results on 3DPW validation set. 'exclude nearby' samples positive images except the ones nearby the anchor. On the other hand, 'include nearby' samples the nearby images. 'skip step' defines the length of nearby frames. If set to 3, previous and next 3 frames are picked. w is an weighting factor of metric learning loss. Each evaluation results are the best performance after running the same condition multiple times.

Method	positive sampling	w	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
VIBE	-	-	**93.587**	56.5637	113.4135	27.1271
Metric	exclude nearby	0.1	94.8115	56.8036	114.71	25.4694
Metric	exclude nearby	0.5	97.5705	**56.4835**	114.8364	**23.4336**
Metric	include nearby	1.0	94.0630	56.5464	113.6545	24.5778
Seg	-	-	96.2915	57.5435	**111.8364**	25.8538

Table 2: Evaluation results on 3DPW test set. w is an weighting factor of metric learning loss. Each evaluation results are the best performance after running the same condition multiple times.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Method	positive sampling	w	MPJPE ↓	PA-MPJPE ↓	Accel ↓
VIBE(paper)	-	-	**97.7**	**63.4**	-
VIBE	-	-	99.1523	64.8514	31.4988
Metric	exclude nearby	0.1	100.0493	65.4618	30.8791
Metric	exclude nearby	0.5	100.2639	65.7869	**29.2455**
Metric	include nearby	1.0	102.7147	66.3104	29.9518
Seg	-	-	98.3965	64.7384	30.3696

Table 3: Evaluation results on MPI-INF-3DHP test set. w is an weighting factor of metric learning loss. Each evaluation results are the best performance after running the same condition multiple times. Note that the performance stated in the VIBE paper and our evaluation result on the baseline VIBE does not match.

Table 2 summarizes the evaluation result on 3DPW test set. While MPJPE and PVE are sacrificed, our method achieves better performance on acceleration error and PA-MPJPE metrics. Table 3 shows the evaluation result on MPI-INF-3DHP test set. Since ground-truth SMPL parameters are not provided in MPI-INF-3DHP test set, we do not include PVE metric. We report both the values from the VIBE paper and our own evaluation results, due to reproducibility issue. Similarly, our method consistently is superior in terms of acceleration error. From the evaluation results on 3DPW test dataset, including nearby frames was second to the VIBE in terms of MPJPE and PVE, while PA-MPJPE and acceleration error did not outperform the 'include nearby' sampling strategy.

To summarize, our method sacrifices MPJPE and achieves better temporal consistence, as the acceleration error is improved. This shows that our metric loss effectively regularizes the temporal pose generator to produce temporally consistent results, while distinguishing similar and dissimilar poses in the feature space. 'include nearby' positive sampling method was proven to be effective, but its performance was inferior to 'exclude nearby' method on PA-MPJPE and acceleration error. Finding a way to merge these two sampling strategies would be an interesting work.

4.4 Evaluation Results: Segmentation Module

First, we present the generated pseudo-ground-truths for segmentation module training. Figure 2 and Figure 3 visualizes our generated segmentation ground-truths. It shows that our method is reliable to be used for training.

Since the number of pixels labeled as background dominates that of the person(foreground), applying proper class weight would be crucial for segmentation task. We tested with two different class weight configurations. In order to quickly compare the results, we break the iteration after every 1000th step of every epoch. If the class weights are defined as $w_{background}, w_{foreground}$, we set the following two options:

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

- option 1: $w_{background} = 1.0, w_{foreground} = 0.5$
- option 2: $w_{background} = 0.5, w_{foreground} = 1.0$

Figure 4 comparatively illustrates the options. Model trained with the option 1 produces irregular outline of human body, since estimating background became the higher priority. On the other hand, the model trained with the option 2 produced more smooth and continuous foreground area. Therefore, we used the option 2 as our default configuration.

Figure 5 and Figure 6 shows the inference result from the pretrained segmentation module. Note that we didn't train the segmentation module on 3DPW dataset and still performs plausible inference results. This shows that our segmentation module can generalize to unseen domain. Although it is well generalized to the outdoor wild images, it produces spurious results on the background, although we performed noise filtering with threshold value. As illustrated on the middle image of the Figure 6, if the person interacts with an object(e.g. clothes), the segmentation module tends to pay attention on the object. Although it would be helpful when we consider person-object interaction, its negative effect would be far more significant, since the model may map the entire background if the background is too noisy or crowded.

From the results on Table 1, 2, 3, segmentation module showed better performance in terms of PVE compared to the baseline VIBE. Although MPJPE and PA-MPJPE is degraded, our findings suggest that further advancement would be possible with a few engineering efforts.

Unfortunately, we could not further conduct experiments due to the fault during the data preparation. However, our preliminary results signifies that our initiatives and approaches could make more advancement in the future.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

4.5 Evaluation Results: Few-Shot Domain Adaptation

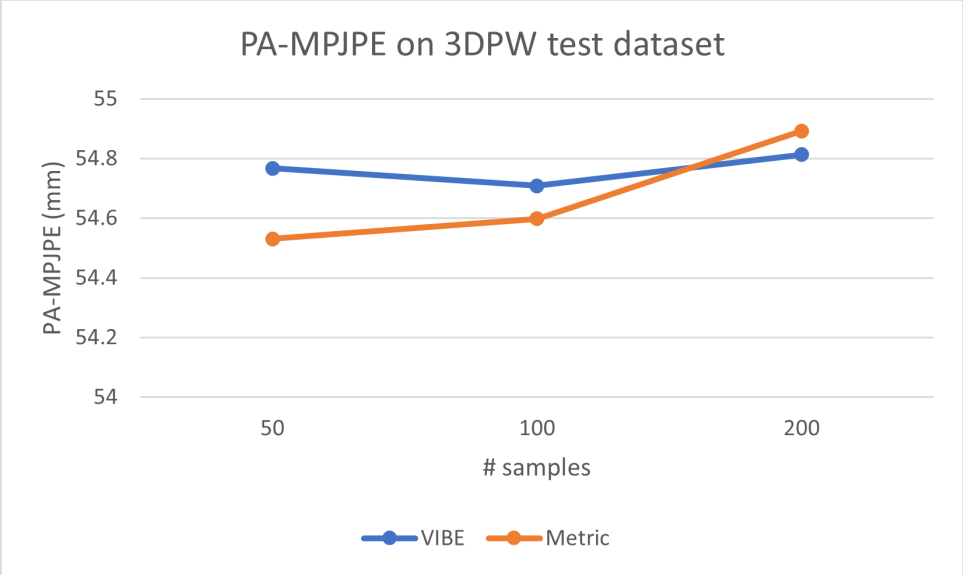


Figure 7: Few-shot evaluation results on 3DPW test dataset. **VIBE** is the baseline VIBE model. **Metric** is our proposed method.

Method	# samples	w	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
VIBE	50	-	**90.7470**	54.7672	107.3041	**26.3925**
Metric	50	1.0	91.0386	**54.5306**	**106.8656**	26.9931

Table 4: Evaluation results on 3DPW test set. w is an weighting factor of metric learning loss. Each evaluation results are the best performance after running the same condition multiple times.

Figure 7 shows the PA-MPJPE scores on 3DPW test dataset under few-shot setting. Although our method was inferior at 200 shots, our method outperformed the baseline model on fewer samples, which can be more effective in many real-world applications. Table 4 reports the details of our few-shot evaluation results. Baseline was better than our method in terms of MPJPE and acceleration error, while our method outperformed the baseline in terms of PA-MPJPE and PVE. One of the possible reasons of inferior performance could be due to adverse effect of feature-level constraint. Applying supervision on SMPL parameter would be better, but its angular representation makes the formulation of supervision far more difficult. Applying constraint on SMPL parameter would be an interesting future work. While their performance difference is insignificant, it shows that our method might have possibility to improve the performance on few-shot test environment.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

4.6 Qualitative Analysis

Figure 8 and Figure 9 illustrates the validity of our metric loss method compared to VIBE. Both figures show that our method more accurately estimates the orientation and positions of the feet, while producing temporally smooth results. Although it is not obvious, VIBE's prediction jitters between frames, especially on the limb parts. Our method, however, relatively produces more smooth translation of the body parts, producing more natural human motion.

Figure 10 shows failure cases from 3DPW data. In case of the left image, subjects are in the car, can the subjects cannot be identified. It seems rather abnormal to correctly localize the subjects. However, if we apply temporal framework onto segmentation network so that the network predicts the foreground area from the previous frames, we might have different result. It would be another interesting future work. In case of the right image, on the other hand, the model failed to ignore the unrelated area, so high probabilities are mapped nearby the on the ground. lighting and color issues make the model hard to differentiate between the subject and the background.

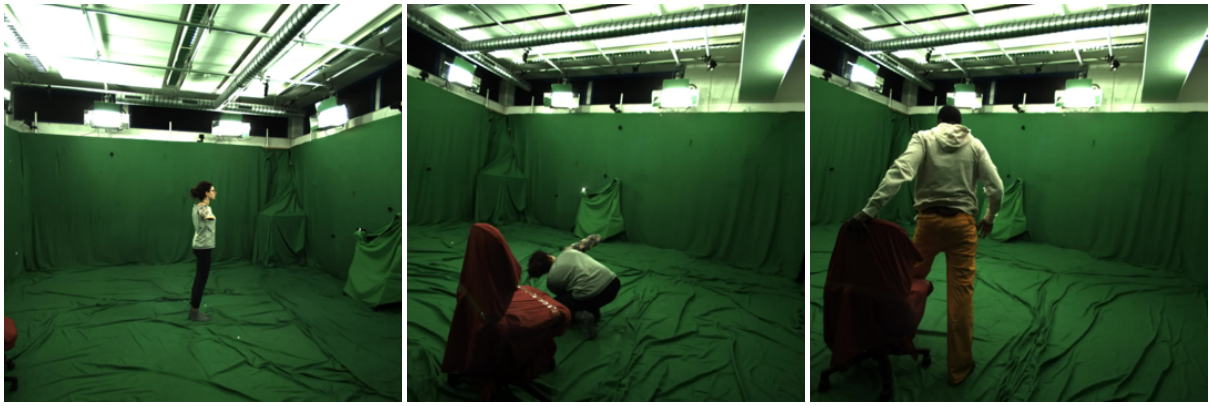
ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

V Conclusion

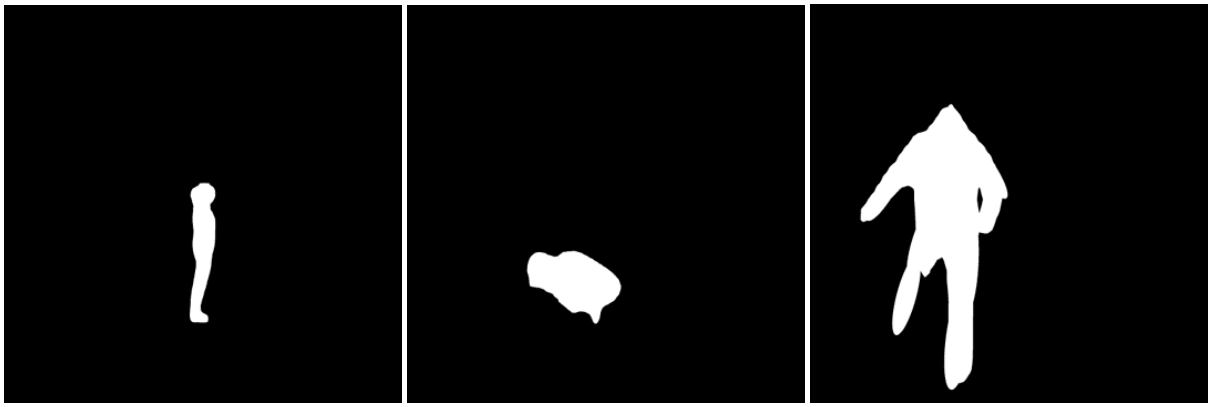
This paper presented metric loss framework with in-batch triplet sampling strategy and foreground segmentation backbone network to produce more robust estimation on unseen test data. We showed that our person segmentation module generalizes well to the unseen wild images by performing evaluation on 3DPW dataset. Although our method is found to be inferior on joint position accuracy, our method can regress more smooth human motion across time. Also, our metric loss showed promising result on few-shot test on 3DPW dataset, which implies that our intuition can be helpful for domain adaptation.

These experimental results incurs promising future works, such as elaborating sampling strategy or applying several different temporal constraints. Also, we can use the encoding part of the pretrained segmentation network for feature extraction, which can significantly reduce the overhead during the feature fusion process. Lastly, making the entire network end-to-end trainable would be an interesting approach, since freezing the weights of the backbone network would not significantly increase the estimation performance.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY



(a) Original MPI-INF-3DHP images

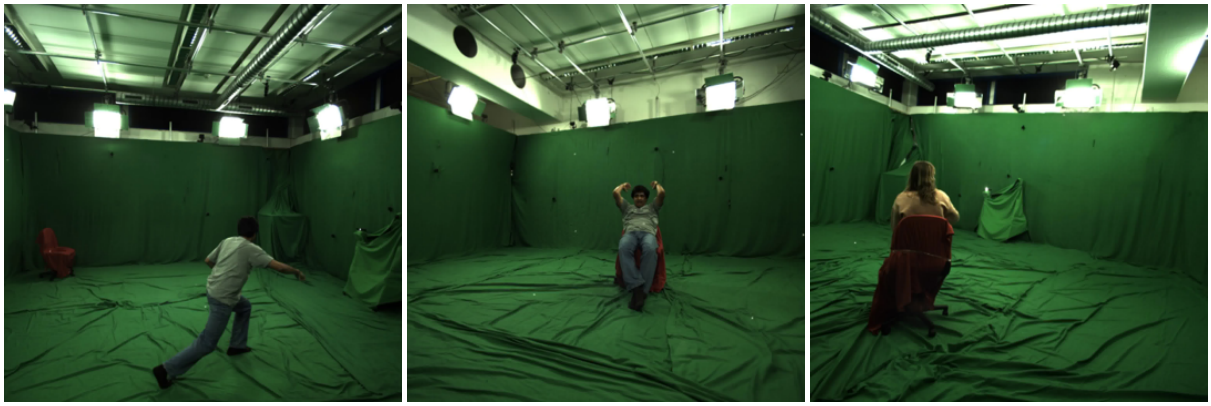


(b) Pseudo-ground-truth segmentation mask

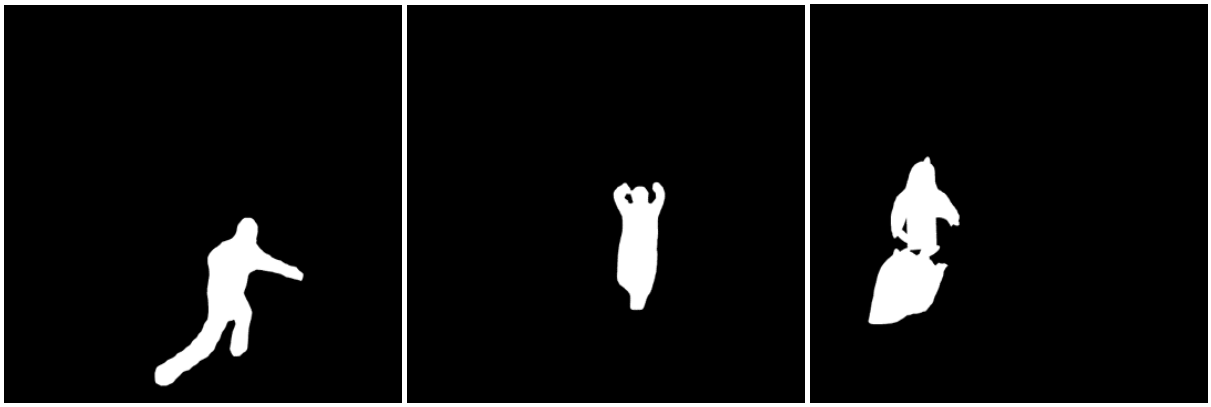
(c) Interpolation of original image and pseudo-ground-truth segmentation mask

Figure 2: Visualization of pseudo-ground-truth segmentation mask generation process. Topmost row shows the original MPI-INF-3DHP training images. Middle row is the generated segmentation mask. Bottom row shoes the interpolation of the original image and the segmentation mask.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY



(a) Original MPI-INF-3DHP images

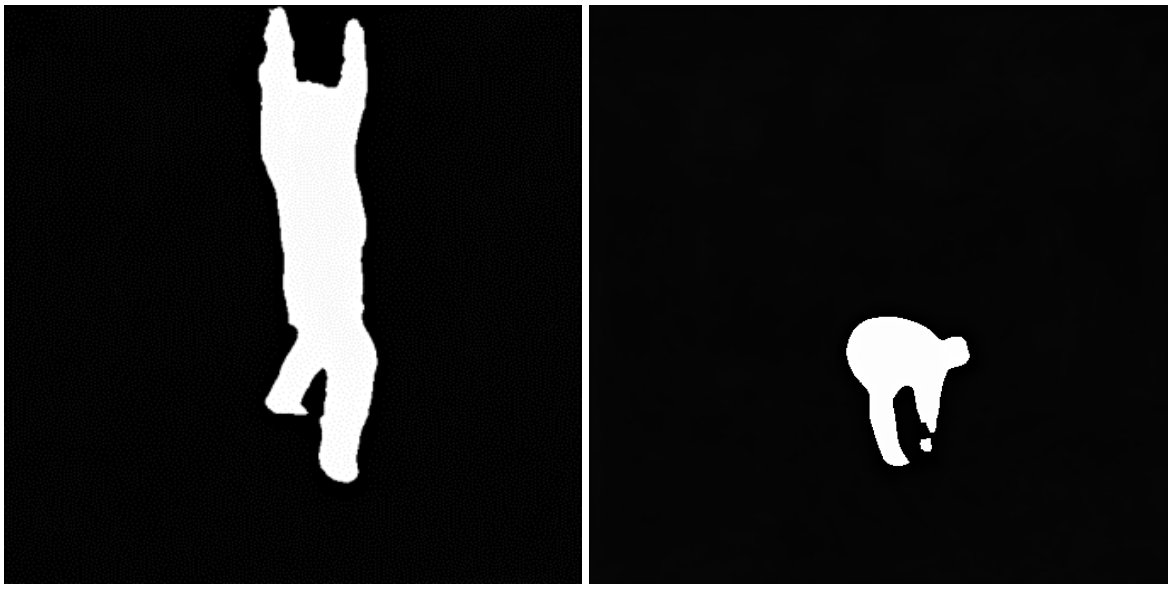


(b) Pseudo-ground-truth segmentation mask

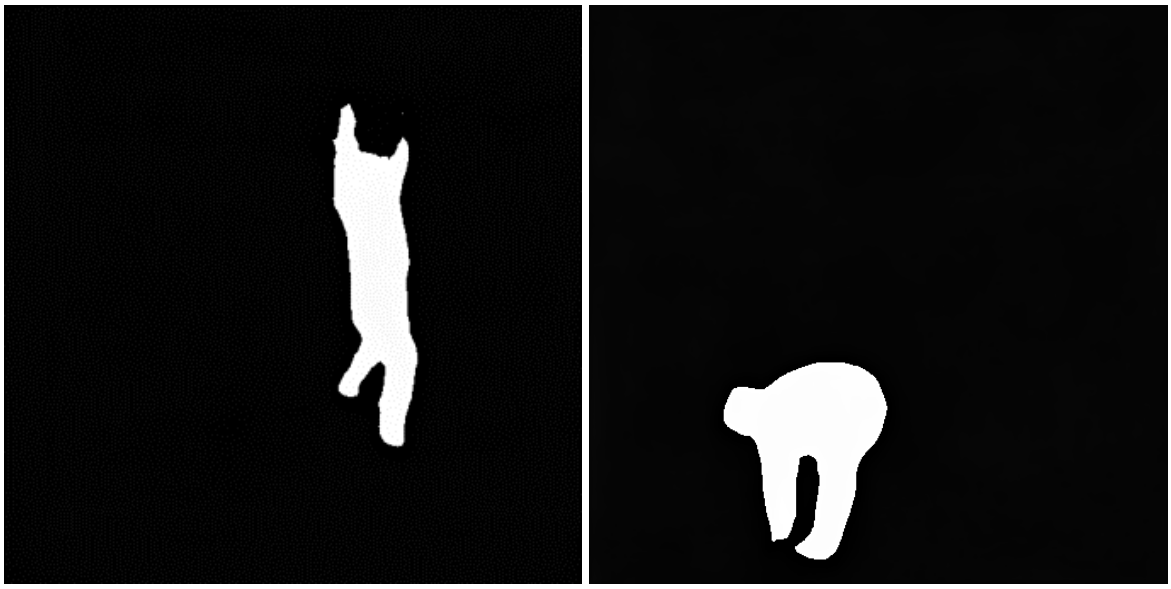
(c) Interpolation of original image and pseudo-ground-truth segmentation mask

Figure 3: Visualization of pseudo-ground-truth segmentation mask generation process. Topmost row shows the original MPI-INF-3DHP training images. Middle row is the generated segmentation mask. Bottom row shoes the interpolation of the original image and the segmentation mask.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY



(a) Option 1



(b) Option 2

Figure 4: Ablation study on MPI-INF-3DHP dataset. Option 1 applies 0.5 on foreground and 1.0 on background class. Option 2 applies 1.0 on foreground and 0.5 on background class

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

(a) MPI-INF-3DHP test images

(b) Inferred segmentation mask

Figure 5: Inference on MPI-INF-3DHP test dataset. Upper row represents the input MPI-INF-3DHP test dataset images. Lower row illustrates inference results.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

(a) MPI-INF-3DHP test images

(b) Inferred segmentation mask

Figure 6: Inference on 3DPW dataset. Upper row represents the input 3DPW dataset images. Lower row illustrates inference results. Note that we did not perform any training on 3DPW dataset.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

(a) VIBE

(b) Our method

Figure 8: Visualization result of our method on 3DPW downtown_sitOnStairs data.

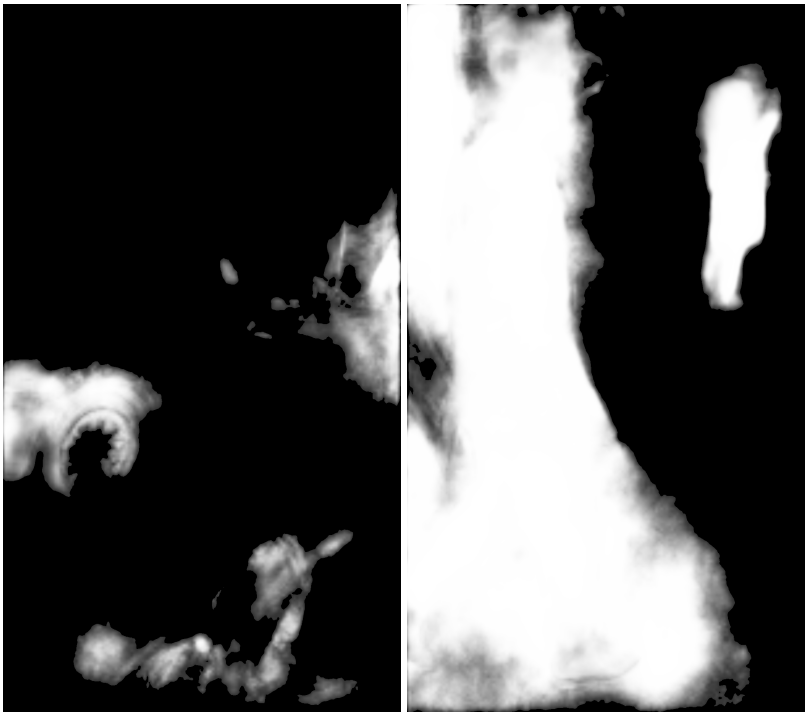
(a) VIBE

(b) Our method

Figure 9: Visualization result of our method on 3DPW downtown_walking data.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

(a) 3DPW images



(b) Segmentation module inference results

Figure 10: Failure cases from segmentation network inference.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

References

[1] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *International Conference on Computer Vision (ICCV)*, 2019.

[2] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," 2018.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[4] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[5] R. Xie, C. Wang, and Y. Wang, "Metafuse: A pre-trained fusion model for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain, "Multiview-consistent semi-supervised learning for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Z. Zhang, C. Wang, W. Qin, and W. Zeng, "Fusing wearable imus with multi-view images for human pose estimation: A geometric approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] H. Tu, C. Wang, and W. Zeng, "Voxelpose: Towards multi-camera 3d human pose estimation in wild environment," in *European Conference on Computer Vision (ECCV)*, 2020.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[10] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Regognition (CVPR)*, 2018.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[11] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017. [Online]. Available: http://up.is.tuebingen.mpg.de

[12] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019.

[13] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[15] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "Hemlets posh: Learning part-centric heatmap triplets for 3d human pose and shape estimation," 2021.

[16] Z. Luo, S. A. Golestaneh, and K. M. Kitani, "3d human motion estimation via motion compression and refinement," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[17] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecká, and Z. Wu, *Hierarchical Kinematic Human Mesh Recovery*, 11 2020, pp. 768–784.

[18] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang, "Sequential 3d human pose and shape estimation from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," *CoRR*, vol. abs/2012.09760, 2020. [Online]. Available: https://arxiv.org/abs/2012.09760

[20] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," 2021.

[21] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] W. Liu, J. Chen, C. Li, C. Qian, X. Chu, and X. Hu, "A cascaded inception of inception network with attention modulated feature fusion for human pose estimation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*,

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 7170–7177. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17206

[23] R. Liu, J. Shen, H. Wang, C. Chen, S.-c. Cheung, and V. Asari, "Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[25] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," 2015.

[28] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018.

[29] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] W. Zhang and C. Xiao, "Pcan: 3d attention map learning using contextual information for point cloud based retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[33] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[34] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[35] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[36] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10 631–10 638, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6689

[37] S. Zhao, Y. Sheng, Y. Dong, E. I.-C. Chang, and Y. Xu, "Maskflownet: Asymmetric feature matching with learnable occlusion mask," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[38] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 611–625.

[39] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Z. Jiang, B. Liu, S. Schulter, Z. Wang, and M. Chandraker, "Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] A. Kortylewski, J. He, Q. Liu, , and A. Yuille, "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[42] J. Piao, C. Qian, and H. Li, "Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9397–9406.

[43] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[44] L. Tang, D. Wertheimer, and B. Hariharan, "Revisiting pose-normalization for fine-grained few-shot recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[45] J. Cao, H. Tang, H.-S. Fang, X. Shen, C. Lu, and Y.-W. Tai, "Cross-domain adaptation for animal pose estimation," 2019.

[46] C. Yang and S.-N. Lim, "One-shot domain adaptation for face generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[47] S. Ling, A. Pastor, J. Li, Z. Che, J. Wang, J. Kim, and P. L. Callet, "Few-shot pill recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[48] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved few-shot visual classification," 2020.

[49] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, and C. Achard, "Pandanet: Anchor-based single-shot multi-person 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[50] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Perez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[51] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1.

[52] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682

[53] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf

[54] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5022–5030.

[55] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2288–2297.

[56] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[57] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2019. [Online]. Available: http://dx.doi.org/10.1109/tpami.2019.2956516

[58] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. [Online]. Available: http://gvv.mpi-inf.mpg.de/3dhp_dataset

[59] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *European Conference on Computer Vision (ECCV)*, sep 2018.

[60] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.

[61] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "Amass: Archive of motion capture as surface shapes," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5441–5450.

[62] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2021.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Acknowledgements

I thank professor Kwang In Kim at UNIST for his support during my master's course. I would have been impossible to get here without his support. I also thank professor Seungryul Baek at UNIST and professor Hyung Jin Chang at University of Birmingham for helpful advice.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY